ALTERNATIF MODEL REGRESI *GROSS NATIONAL INCOME* NEGARA-NEGARA ASIA TERHADAP HARAPAN HIDUP

(Studi Hasil Regresi Data Frank J. Anscombe)

Mochamad Setyo Pramono[1]

ABSTRACT

Regression analysis is used to model relationships between variables and to determine the magnitude of the relationships. Then the models can be used to make predictions. Franc Anscombe, statistician gave case examples of regression analysis, which hyphothetics data were in the forms of some independent variables and dependent variables. Interesting phenomenon from Anscombe's data were the result of data processing would produce the same statistical values although the data were different one from the other. Pursuant from regression resuls in the study of Frank Anscombe's data, this research compared the correct regression models between life expectancy with Gross National Income (GNI) among some countries in Asia. Results of scatter plots showed that the data pattern was not be obvious, if tending to be linear or quadratic. Therefore, both approaches were used in regresision analysis to be compared. Analysis results showed that the linear regression model for Live Expectancy is = 63.46 + 0.00069 GNI with R^2 = 68.27% and the quadratic regression model for Live Expectancy is = 60.62 + 0.0015 GNI – 2.94E-08 GNI^2 with R^2 = 73.51%. If R^2 as determinant to choose the best model, so the model of quadratic regression is choosen because it has bigger R^2 than the R^2 in the linear regression.

Key words: *Regression analysis, Frank Anscombe, life expectancy, Gross National Income*

PENDAHULUAN

Istilah regresi pertama kali justru diperkenalkan oleh seorang antropolog dan ahli meteorologi dari Inggris, Francis Galton (1822–1911) pada penelitian tentang sifat-sifat keturunan dan masalah biologi. Selama ini analisis regresi merupakan salah satu dari metode statistika yang cukup populer untuk menganalisis data. Proses analisis data pada dasarnya meliputi upaya penelusuran dan pengungkapan informasi yang relevan yang terkandung dalam data. Pada tahap awal, suatu analisis data diusahakan tanpa terlalu terikat pada asumsi-asumsi yang ketat agar pengungkapan informasi dapat dilakukan dengan fleksibel dan lebih merangsang imajinasi tanpa melupakan kaidah-kaidah teori yang dikenal (Aunuddin, 1989). Hal ini dimungkinkan karena kita dihadapkan pada hal-hal yang tak terduga yang mungkin jauh lebih menarik dibandingkan persoalan semula.

Kecenderungan yang kurang tepat terjadi dalam penerapan metode statistika secara umum, bahwa penelitian yang sebenarnya dapat dibahas lebih menarik ternyata cukup diakhiri dengan kesimpulan yang menyatakan hasilnya bermakna (signifikan) atau tidak bermakna secara statistik. Ilustrasi Tabel 1 mungkin dapat dijadikan contoh output analisis regresi linier sederhana dari suatu data dengan variabel bebas X dan variabel tak bebas Y2.

Diperlukan sedikit 'pengetahuan' tentang analisis statistik terutama persamaan regresi untuk membaca output di atas. Bila diperhatikan model persamaan regresi di atas **Y2 = 3,00 + 0,500X**, sudah bermakna beserta dengan dugaan koefisien regresinya. Kesimpulan ini diambil di mana salah satu cara yang paling mudah dengan membandingkan *p-value* dengan tingkat signifikansi α = 0,05. Jika *p-value* < α, maka persamaan regresi atau koefisien regresi dianggap cukup signifikan. Sementara itu koefisien

[1] Pusat Penelitian dan Pengembangan Sistem dan Kebijakan Kesehatan, Jl. Indrapura No. 17, Surabaya 60176
Korespondensi: Mochammad Setyo Pramono
Pusat Penelitian dan Pengembangan Sistem dan Kebijakan Kesehatan
Jl. Indrapura No. 17, Surabaya 60176
Email: yoyokpram@yahoo.com

Tabel 1. Output regresi linier antara variabel Y2 dan X

Regression Analysis: Y2 versus X
The regression equation is
Y2 = 3.00 + 0.500 X

Predictor	Coef	SE Coef	T	P
Constant	3.001	1.125	2.67	0.026
X	0.5000	0.1180	4.24	0.002

S = 1.237 R-Sq = 66.6% R-Sq(adj) = 62.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	27.500	27.500	17.97	0.002
Residual Error	9	13.776	1.531		
Total	10	41.276			

determinasi (R^2) = 66,6% menunjukkan prosentase kedekatan antara titik-titik pengamatan dengan garis regresinya yang dalam bahasa sederhana adalah tingkat akurasi model regresi. Istilah lain dari R^2 adalah koefisien korelasi ganda yang digunakan untuk mengukur proporsi keragaman total yang dapat dijelaskan dalam data (Drapper dan Smith, 1981). Bila R^2 = 100%, berarti semua titik-titik pengamatan akan tepat berada pada garis regresinya.

Tidak ada yang salah dalam hasil di atas, hanya pertanyaannya adalah apakah hasil analisis di atas dapat menunjukkan suatu pola tertentu dan apakah hasil tersebut memerlukan penjelasan lebih lanjut, atau dengan kata lain apakah model di atas sudah mewakili (pola) datanya? Apakah tidak ada model lain yang lebih baik?

Berangkat dari pertanyaan ini maka tulisan ini bertujuan untuk mencari model regresi terbaik melalui eksplorasi data dengan pendekatan *scatter plot* dan membandingkan hasil output analisis regresi melalui studi kasus data *Anscombe* serta selanjutnya mencoba menerapkannya pada regresi antara *Gross National Income* (GNI) dan harapan hidup negara-negara di Asia.

METODE

Penelitian ini menggunakan data sekunder dengan design potong lintang data *Gross Nacional Income* (GNI) dan harapan hidup negara-negara di Asia tahun 2004 yang berasal dari *Population Reference Bureau* (PRB). Sebagai studi kasus untuk bantuan analisis digunakan data hipotetik dari *Frank Anscombe*. Analisis yang digunakan adalah regresi sederhana yang dibantu dengan *scatter plot*.

HASIL DAN PEMBAHASAN

Data Frank J. Anscombe

Frank J. Anscombe, seorang ahli statistika lulusan Cambridge University termasuk ilmuwan yang peduli pada masalah-masalah lingkungan dan kesehatan. Pada jurnal *The American Statistician* edisi bulan Pebruari 1973, *Anscombe* menulis yang dalamnya memberikan contoh menarik yaitu berupa data hipotetik yang berupa variabel bebas (X) dan variabel tak bebas (Y) yang dianalisis secara berpasang-pasang seperti disajikan pada Tabel 2.

Tabel 2. Data Frank Anscombe

Data 1		Data 2		Data 3		Data 4	
X	Y1	X	Y2	X	Y3	X4	Y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Fenomena yang menarik dari data di atas adalah hasil pengolahan data tersebut akan menghasilkan besaran-besaran statistik yang identik, mulai dari rata-ratanya, koefisien korelasi (besaran hubungan antara X dan Y), persamaan (model) garis regresi, hasil uji anova, sampai dengan R^2 memiliki nilai yang sama. Contoh output analisis regresi di awal tulisan ini (Tabel 1) merupakan hasil proses dari data *Anscombe*. Rekapitulasi hasil olahan statistik dari masing-masing kelompok data *Anscombe* disajikan pada Tabel 3.

Jika di dunia jurnalistik dikenal istilah "sebuah gambar itu memuat seribu kata", analog dengan itu sebuah grafik dapat mencerminkan seratus statistik. Istilah lain yang memiliki makna sama adalah *speak*

Tabel 3. Statistik masing-masing kelompok data *Anscombe*

Statistik	Data 1 (X, Y1)	Data 2 (X, Y2)	Data 3 (X, Y3)	Data 4 (X4 , Y4)
Rata-rata X	9,00	9,00	9,00	9,00
Rata-rata Y	7,50	7,50	7,50	7,50
Garis regresi	Y1 = 3 + 0,5X	Y2 = 3 + 0,5X	Y3 = 3 + 0,5X	Y4 = 3 + 0,5X4
Koefisien korelasi	0,82	0,82	0,82	0,82
R^2	66,7%	66,6%	66,6%	66,7%

with data atau *you must look at data* (Davies, 1998). Mencermati hal tersebut maka jika data *Anscombe* ini kita plot, hasilnya menunjukkan pola yang sama sekali berbeda satu dengan yang lain sebagaimana pada Gambar 1.

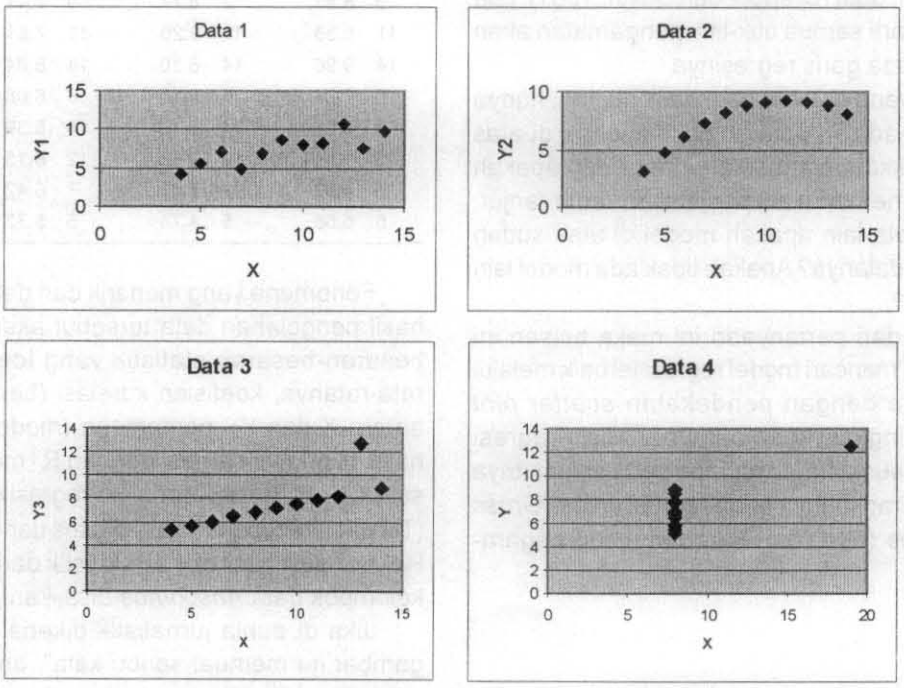
Adalah menarik, dengan pola data yang berbeda satu dengan yang lainnya ternyata menghasilkan model regresi yang sama persis. Pertanyaannya, apakah model-model tersebut sudah mewakili pola datanya?

Data 1

Gambar 1 menunjukkan bentuk pola plot data 1 memiliki tren cenderung naik dengan komposisi jarak yang agak melebar antara amatan satu dengan yang lain. Tampaknya regresi linier sederhana merupakan solusi yang sesuai dengan kondisi data tersebut.

Data 2

Sedangkan pola plot untuk data 2 menunjukkan bentuk tidak linier atau lurus, tetapi melengkung, maka pendekatan model regresi yang ideal dengan cara menggunakan estimasi bentuk kurva dan bukan langsung diolah dengan regresi linier. Sebagaimana diketahui, bila menggunakan pendekatan model regresi kuadratik maka untuk data 2 diperoleh persamaan: $Y2 = -559{,}573 + 278{,}084\ X - 12\ X^2$ dengan R^2 sebesar 100%.



Gambar 1. *Scatter plot* data *Frank J. Anscombe*

Data 3

Bila diamati plot untuk data 3 yaitu pola linier tetapi ada satu masalah yaitu ada amatan (data) yang memencil, pada koordinat: X (X = 13, Y = 12.74). Kita harus berhati-hati pada amatan ini, apakah dia berupa pencilan (*outlier*) atau amatan yang berpengaruh (*influence*). Jika dia *outlier* harus ditelaah terlebih dahulu apakah ada kesalahan pada proses pengambilan data atau tidak. *Outlier* merupakan suatu keganjilan dan menandakan suatu titik data yang sama sekali tidak tipikal dibandingkan data yang lain (Hair, *et al.* 1995). Pembuangan *outlier* hanya dapat dilakukan jika memang diketahui dengan pasti ada kesalahan baik dalam pengukuran atau pencatatan. Baik *outlier* maupun *influence* memiliki karakteristik yang berbeda walaupun kedua gejala ini disebabkan oleh pengamatan yang sama.

Data 4

Sedangkan pola dari data 4 sebenarnya tidak dapat dilakukan pendekatan dengan persamaan regresi. Hal ini disebabkan variabel independen (X) konstan sedangkan variabel dependen (Y) yang dihasilkan berbeda. Tetapi memang ada satu amatan yang memencil pada koordinat (X = 19, Y = 12.50). Jika amatan ini dihilangkan maka persamaan garisnya hanya berupa X = konstanta. Dengan kata lain, amatan ini "sangat berpengaruh" karena tanpa satu titik (amatan) ini sebetulnya dapat dikatakan tidak ada hubungan sama sekali antara variabel X dengan Y.

Dari hasil dan pembahasan data *Anscombe* ini, maka akan dicari bagaimana pola hubungan antara harapan hidup dan GNI negara-negara di Asia.

Hubungan antara Harapan Hidup dan *Gross National Income* Negara-Negara Asia.

Hampir setiap tahun *Population Reference Bureau* (PRB) yang berdiri sejak 1929, mengeluarkan data populasi, kesehatan dan lingkungan negara-negara hampir di seluruh dunia, termasuk data harapan hidup dan GNI. Menurut PRB yang dimaksud harapan hidup disini (*life expectancy at birth*) adalah rata-rata usia bayi baru lahir dapat diharapkan terus hidup sampai tingkat kematian tertentu. Dalam bahasa yang lebih populer harapan hidup adalah usia rata-rata masyarakat di suatu negara. Ada banyak indikator, baik langsung maupun tidak langsung yang diduga mempengaruhi harapan hidup masyarakat, antara lain status kesehatan, tingkat pendidikan, kemampuan ekonomi sampai dengan kondisi sosial politik. Negara yang terancam perang saudara memiliki harapan hidup yang rendah misalnya di Rwanda, Afrika, harapan hidupnya tidak lebih dari 44 tahun.

Variabel yang dianggap mewakili indikator-indikator tersebut adalah GNI per kapita suatu negara. Menurut PRB, yang dimaksud GNI disini adalah GNI *Purchasing Power Parity* atau GNI PPP per kapita dibagi populasi tengah tahun. Negara yang menurut tingkat ekonominya mapan, hampir bisa dipastikan memiliki kondisi sosial politik, pendidikan sampai dengan tingkat kesehatan masyarakat yang mapan pula. Bagaimana hubungan antara harapan hidup dengan GNI di suatu negara. Data yang ada menunjukkan bahwa negara-negara yang maju dengan GNI yang tinggi cenderung memiliki usia harapan hidup pada masyarakatnya tinggi. Sebaliknya negara-negara yang tertinggal dengan GNI yang rendah cenderung usia harapan hidupnya juga rendah. Untuk pembahasan ini, sebagai batasan data dipilih dari beberapa negara di wilayah Asia sebagaimana disajikan di Tabel 6.

Besaran korelasi antara harapan hidup dengan GNI akan diperoleh nilai sebesar 0,826. Besaran angka ini menunjukkan hubungan yang cukup erat antar kedua variabel tersebut, akan tetapi belum menunjukkan bagaimana polanya. Untuk mengetahui pola hubungannya maka harus di buat persamaan regresinya terlebih dahulu. Pertanyaannya adalah model regresi seperti apa yang tepat? Teori konfirmatori menjelaskan bahwa GNI yang mempengaruhi harapan hidup dan bukan sebaliknya, dengan demikian GNI menjadi variabel bebas (*prediktor*) sedangkan yang menjadi variabel tidak bebas (*respon*) adalah harapan hidup.

Sebagai langkah awal kita plot terlebih dahulu datanya sehingga akan terlihat pola yang nantinya menjadi dasar dalam membuat suatu model regresi sederhana. Ternyata pola yang didapatkan tidak terlihat secara jelas apakah linier atau kuadratik (Gambar 2) sehingga perlu dicobakan kedua pendekatan ini, regresi secara linier dan kuadratik, untuk membandingkan mana yang memberikan hasil terbaik.

Tabel 6. Harapan hidup dan GNI beberapa negara di Asia

Negara	Harapan Hidup (tahun)	GNI ($)	Negara	Harapan hidup (tahun)	GNI ($)
Kamboja	57	1.970	Banglades	60	1.770
Indonesia	68	3.070	India	62	2.650
Laos	54	1.660	Iran	69	6.690
Malaysia	73	8.500	Hong Kong	81	27.480
Philipina	70	4.450	Macao	77	21.910
Singapura	79	23.730	Bahrain	74	16.190
Thailand	71	6.890	Israel	79	19.000
Vietnam	72	2.300	Kuwait	78	17.780
China	71	4.520	Oman	74	13.000
Jepang	82	27.380	Arab Saudi	72	12.660
Korsel	77	16.960	UEA	74	24.030

Sumber: 2004 *World Population Data Sheet of The Population Reference Bureau*

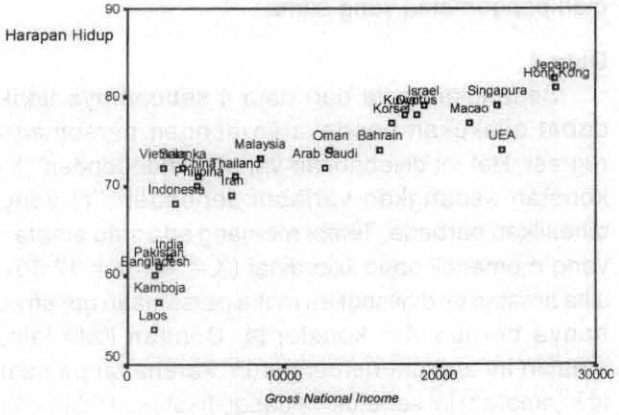
Tabel 7. Output regresi linier antara harapan hidup dengan Gross National Income

```
Dependent variable.. LIFEEXP Method.. LINEAR

R Square .68269
 Analysis of Variance:
 DF Sum of Squares Mean Square

Regression 1 921.63139 921.63139
Residuals 23 428.36861 18.62472
F = 49.48430 Signif F = .0000

---------- Variables in the Equation ----------

Variable B SE B Beta T Sig T
GNI .000688 9.7821E-05 .826251 7.035 .0000
(Constant) 63.455763 1.421394 44.643 .0000
```

Dari tabel 8 menunjukkan hasil jika menggunakan pendekatan regresi linier model adalah Harapan Hidup = 63.46 + 0.00069 GNI dengan R^2 = 68,27%.

Sebaliknya jika menggunakan pendekatan regresi kuadratik sebagaimana disajikan pada Tabel 9 maka di dapat model: Harapan Hidup = 60,62 + 0,0015 GNI – 2,94E-08 GNI^2 dengan R^2 = 73,51%.

Dengan menggunakan α = 0,05, melalui uji ketepatan model yang sama-sama signifikan berikut koefisien regresinya, maka untuk kasus GNI terhadap harapan hidup di negara-negara Asia, persamaan model regresi kuadratik yang menjadi pilihan karena memiliki R^2 yang lebih besar dibandingkan model regresi linier.



Gambar 2. Plot harapan hidup dan GNI di beberapa negara di Asia

KESIMPULAN DAN SARAN

Kesimpulan

Baik model pertama (linier) maupun kedua (kuadratik) memiliki interpretasi umum yang sama, yaitu semakin tinggi GNI suatu negara maka usia harapan hidupnya juga cenderung semakin panjang. Interpretasi secara matematis dapat dilihat dari bentuk dan koefisien persamaannya. Dari hasil *scatter plot* mungkin sebagian orang beranggapan atau dapat menerima model yang pertama karena secara visual tidak begitu tegas polanya apakah linier atau kuadratik. Akan tetapi kalau kita telaah lebih jauh dengan pendekatan non statistik, hubungan antara harapan hidup dengan GNI tidak akan bisa linier. Hal ini dikarenakan usia sebagai ukuran dari harapan hidup cenderung memiliki batas maksimal, sementara

Tabel 8. Output regresi kuadratik antara harapan hidup dengan *Gross National Income*

Dependent variable.. LIFEEXP Method.. QUADRATI

R Square .73509

Analysis of Variance:

	DF	Sum of Squares	Mean Square
Regression	2	992.37326	496.18663
Residuals	22	357.62674	16.25576

F = 30.52374 Signif F = .0000

---------- Variables in the Equation ----------

Variable	B	SE B	Beta	T	Sig T
GNI	.001471	.000386	1.766525	3.808	.0010
GNI**2	-2.93867408E-08	1.4087E-08	-.967738	-2.086	.0488
(Constant)	60.619845	1.900382		31.899	.0000

GNI cenderung tidak memiliki batas maksimal. Dengan kata lain model linier tersebut kurang valid jika dipakai untuk prediksi. Kesimpulannya baik dengan pendekatan statistik maupun non statistik, model regresi kuadratik adalah lebih fisibel di mana modelnya adalah Harapan Hidup = 60,62 + 0,0015 GNI – 2,94E-08 GNI^2 dengan R^2 = 73,51%.

Saran

Pendekatan *scatter plot* yang merupakan cara sederhana untuk pemeriksaan data tetapi ini hanya bisa diterapkan jika variabel dependen dan independen hanya ada satu. Sebenarnya pembahasan untuk memilih persamaan regresi yang terbaik memerlukan analisis yang lebih jauh terutama jika menyangkut multivariabel. Banyak faktor diduga dapat mempengaruhi harapan hidup. Mungkin jika kita masukkan faktor-faktor lain sebagai variabel independennya akan menghasilkan model yang lebih baik. Hal lainnya adalah pada analisa regresi diasumsikan setiap sisaan (*error*) bersifat IIDN, independen, identik dan berdistribusi normal, yang tidak dibahas dalam penelitian ini. Diperlukan eksplorasi lebih jauh tentang sisaan (*error*) agar data tersebut dapat memberikan informasi lebih banyak.

DAFTAR PUSTAKA

Anscombe, Frank J, 1973. Graphs in Statistical Analysis, *American Statistician*, Vol. 27.

Aunuddin, 1989. *Analisa Data*. Bogor: IPB.

Davies, 1998. The Value of Pictures. *Spectroscopy Europe 10/4/1998.*

Draper NR dan Smith H, 1981. *Analisa Regresi Terapan*, Edisi 2, New York: John Willey.

Hair FJ, Anderson ER, Tatham LR, and Black CW, 1995. *Multivariate Data Analysis with Readings*, Fourth edition, New Jersey: Prentice-Hall.

The Reference Population Bureau, 2004. *World Population Data Sheet.*